Running head: BAYESIAN VARIANCE COMPONENT ESTIMATION

Bayesian Variance Component Estimation Using the Inverse-Gamma Class of Priors in a Nested Generalizability Design

Ethan A. Arenson

University of Connecticut

Paper presented at the annual meeting of the New England Research Association, October 22, 2009, Rocky Hill, Connecticut.

Abstract

One of the problems inherent in variance component estimation centers around inadmissible estimates. Such estimates occur when there is more variability within groups, relative to between groups. This paper suggests a Bayesian approach to resolve inadmissibility by placing noninformative inverse-gamma priors on the variance components, and compares Bayesian estimates with expected mean square and maximum likelihood estimates. No noticeable differences among estimation type were found for balanced data. However, Bayesian estimates tended to produce less biased estimates for unbalanced data.

Bayesian Variance Component Estimation Using the Inverse-Gamma Class of Priors in a Nested Generalizability Design

Assessments are fallible instruments. That is, it is not possible to measure with certainty the latent construct of interest. In educational settings, the construct of interest typically is ability or achievement. Given the need to design highly accurate measurement processes, researchers have given considerable attention to test reliability as a way of estimating accuracy. The most commonly used reliability measures have severe limitations. For example, Kuder and Richardson's formulae 20 and 21 (1937) for dichotomous items, as well as Cronbach's coefficient alpha (1951) are based on classical test theory (CTT; Lord & Novick, 1968; Crocker & Algina, 1986) assumptions. CTT posits that variability of observed scores can be decomposed into two components, namely persons, τ, and error, ϵ:

$$X = \tau + \epsilon \tag{1}$$

Furthermore, KR-20 assumes constant inter-item correlations; KR-21 assumes the items are of equal difficulty; and coefficient alpha is merely a lower bound on the unknown reliability of the test.

Generalizability theory (GT; Cronbach, Rajaratnam, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001) extends the true-score model posited by CTT. GT allows for multiple sources of variability. In educational applications, such additional sources typically include variability due to items, raters, or occasions. For example, one can express a design in which each of P persons responds to I unique items from a universe of equivalent items (often referred to as an $i : p$, or nested, design) as

$$X_{i:p} = \mu + \alpha_p + \epsilon_{i:p} \, . \tag{2}$$

The fundamentals of GT are rooted in variance component models, which are most

commonly found in analysis of variance. One typically estimates variance components by equating ANOVA mean squares to their expectation. This method is known as EMS estimation. Maximum likelihood (ML) estimation, generally considered to be a preferred method, can be shown to be a function of EMS estimates (Searle, Casella, & McCulloch, 2006). Estimation of variance components is discussed below and in the Methods section.

Expected Mean Squares

The expected mean squares (EMS; Cronbach et al., 1972; Brennan, 2001) method equates observed mean squares to their expectations. Table 1 summarizes the expected mean squares for balanced data in the nested model, Equation (2).

Insert Table 1 here.

One sees that when MS_p is less than $\mathrm{MS}_{i:p}$–that is, when there is more variability within groups than there is between groups–the variance component for persons, $\hat{\sigma}^2_p$, is negative. Such results are inadmissible and have serious implications, as addressed below.

Inadmissible Variance Component Estimates

In any ANOVA design, mean squares–and, hence, variance components–are random quantities. As such, there is always some chance of a negative variance component estimate via classical estimation methods. Under normality assumptions, these quantities are independent, and one can use the F-distribution to determine this probability (Searle et al., 2006): $F = \frac{\mathrm{MS}_p}{\mathrm{MS}_{i:p}} \sim F^{\nu_p}_{\nu_e}$, where F^n_d refers to a random variable from the F distribution with n numerator and d denominator degrees of freedom. It follows that

$$\Pr(\hat{\sigma}^2_p < 0) \quad = \quad \Pr\left(F^{\nu_p}_{\nu_d} < 1\right). \tag{3}$$

While Novick, Jackson, and Thayer (1971) described such a situation involving negative variance components as “somewhat absurd,” Brennan (2001) pointed out that this

happens either when the sample size is small or when there are a large number of effects, either of which lead to large values of $\mathrm{MS}_{i:p}$.

Some argue, however, that the presence of negative variance components is characteristic of a more fundamental problem. Frequentist statisticians agree that negative variance component estimates are problematic, suggesting that the data do not support the assumed model (Searle et al., 2006). The frequentist resolution of inadmissability involves setting $\hat{\sigma}_p^2 = 0$. This approach not only results in biased variance component estimates (Brennan, 2001), but it underestimates the magnitude of the bias. In addition, this method of truncated estimation has nothing to do with statistical inference, (Scheffé, 1959).

Bayesian statisticians interpret the negative variance component problem differently. They claim that a negative estimate arises because its likelihood function is uninformative (i.e., flat; Hill, 1965; Tiao & Tan, 1965). Thus, it is not the case that $\sigma_p^2 \approx 0$, but, rather, that the proposed model does not sufficiently explain the data (Nelder, 1954; Hill, 1965; Tiao & Tan, 1965). This explanation seems consistent with that of Brennan, previously mentioned, regarding the number of facets and sample size.

As will be discussed in the Methods section, one can use a Bayesian approach to incorporate prior knowledge (i.e., that variances are positive quantities), to obtain admissible results. The goal in Bayesian analysis is to learn about the posterior distribution, $p(\boldsymbol{\theta}|\text{data})$, which one determines based on combining the likelihood of the data, $L(\text{data}|\boldsymbol{\theta})$, with the prior distribution, $p(\boldsymbol{\theta})$. If $\boldsymbol{\theta}$ denotes a vector of parameters, then the resulting posterior distribution,

$$p(\boldsymbol{\theta}|\text{data}) \propto L(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{4}$$

depends on the choice of prior distribution.

This paper includes priors from the inverse-gamma class. This class of priors holds

many desirable qualities, two of which are presented here. First, this class of priors is conditionally conjugate (Gelman, Carlin, Stern, & Rubin, 2003; Gelman, 2006). That is, the prior and the resulting posterior distributions can both be expressed as inverse-gamma distributions. Second, the inverse-chi-squared distribution, which is often used to model variances, is a special case of the inverse-gamma family.

Method

Data Generation

Balanced Data. 1000 datasets were simulated for each of $P = \{25, 100, 1000\}$ persons crossed with $I = \{10, 50, 100\}$ items, according the nested model outlined in Equation (2):

$$X_{i:p} = \mu + \alpha_p + \epsilon_{i:p}\,, \quad \text{where}$$

$$\alpha_p \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_p), \quad \text{and} \quad \epsilon_{i:p} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{i:p}). \tag{5}$$

For convenience, $\mu \equiv 0$.

When total variance, σ^2_X, is fixed (here $\sigma^2_X = 100$), one can express σ^2_p and $\sigma^2_{i:p}$ in terms of test reliability, ρ (these results are proven in the Appendix):

$$\sigma^2_p = \frac{(IP-1)}{I(P-1) + (IP+P-2)\frac{1-\rho}{\rho}}\,\sigma^2_X\,, \quad \text{and} \tag{6}$$

$$\sigma^2_{i:p} = \frac{(IP-1)}{I(P-1)\frac{\rho}{1-\rho} + (IP+P-2)}\,\sigma^2_X\,. \tag{7}$$

Tables 2 through 4 display the generating variance components, as a function of I, P, ρ, and σ^2_X.

Insert Tables 2 through 4 here.

When the total variance is fixed, variance components also have an upper bound. A fixed total variance consequently fixes the total sum of squares. These upper bounds

correspond to test reliabilities of $\rho = 0$ for $\sigma^2_{i:p}$ and or $\rho = 1$ for σ^2_p:

$$0 \leq \sigma^2_p \leq \frac{PI-1}{I(P-1)}\sigma^2_X, \quad \text{and} \quad 0 \leq \sigma^2_{i:p} \leq \frac{PI-1}{P(I-1)}\sigma^2_X. \tag{8}$$

Unbalanced Data. For the study of unbalanced data, nine cases in which 100 unique, but equivalent, items will be divided among two, three, and four groups, crossed with reliabilities inclusively between 0.1 and 0.9. Tables 5 through 7 outline the proportions among which items will be apportioned and the resulting variance components.

Insert Tables 5 through 7 here.

Because data are unbalanced, there is no uniform measure of test length. As a proxy, a measure of effective test length (Brennan, 2001) will be used. Let $n_{i:p}$ denote the number of items to which person p responds, and let n_T denote the total number of items (which, in this study, equals 100). Then,

$$I_{\text{eff}} = \sum_{p=1}^{P} \frac{n^2_{i:p}}{n_T}. \tag{9}$$

Parameter Estimation

EMS and Non-Truncated EMS Estimation.. Referring back to Table 1, one can solve each EMS equation in terms of the variance components, and show that

$$\hat{\sigma}^2_p = \begin{cases} \frac{\text{MS}_p - \text{MS}_{i:p}}{K} & \text{MS}_p > \text{MS}_{i:p} \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \tag{10}$$

$$\hat{\sigma}^2_{i:p} = \text{MS}_{i:p}. \tag{11}$$

The corrected EMS estimate for variability due to persons will be made using Federer's (1968) non-truncated estimator for balanced data:

$$\tilde{\sigma}^2_p = \hat{\sigma}^2_p + \frac{\text{MS}_{i:p}}{I} e^{-\alpha \text{MS}_p}, \quad 0 < \alpha \leq \frac{1}{\text{MS}_{i:p}}, \tag{12}$$

where $\alpha = 1$ minimizes the bias of $\tilde{\sigma}_p^2$. Since $\frac{1}{\text{MS}_p}$ is likely to be less than one, in particular for samples with low reliability, the choice for α will be $\min(\alpha, 1)$.

Maximum Likelihood Estimation.. Under the nested model, Equation (2), the likelihood of a single observation, $X_{i:p}$ is proportional to

$$\left(\frac{\text{SS}_p}{\sigma_p^2}\right)^{-\frac{P}{2}} \left(\frac{\text{SS}_{i:p}}{\sigma_{i:p}^2 + I\sigma_p^2}\right)^{-\frac{-[P(I-1)+1]}{2}} \times e^{-\frac{1}{2}\left[\frac{\left(X_{i:p}-\mu\right)^2}{\sigma_{i:p}^2+I\sigma_p^2} + \frac{\text{SS}_p}{\sigma_p^2} + \frac{\text{SS}_{i:p}}{\sigma_{i:p}^2}\right]}. \tag{13}$$

In variance component estimation, μ is a nuisance parameter, which one typically integrates out of the likelihood function. Hence, the expression in Equation (13) reduces to

$$\left(\sigma_X^2\right)^{-\frac{1}{2}} \left(\frac{I\text{SS}_p + \text{SS}_{i:p}}{\sigma_p^2}\right)^{-\frac{P}{2}-1} \left(\frac{\text{SS}_{i:p}}{\sigma_{i:p}^2}\right)^{-\frac{P(I-1)}{2}-1} \times e^{\left[-\frac{1}{2}\left(\frac{\text{SS}_p}{I\sigma_p^2+\sigma_{i:p}^2} + \frac{\text{SS}_{i:p}}{2\sigma_{i:p}^2}\right)\right]}. \tag{14}$$

The complete likelihood is the product of the singleton likelihoods, across persons and items. That is,

$$L(\sigma_p^2, \sigma_{i:p}^2|\text{data}) = \prod_{p=1}^{P}\prod_{i=1}^{I} L(\sigma_p^2, \sigma_{i:p}^2|X_{i:p}, \text{SS}_p, \text{SS}_{i:p}, \nu_p, \nu_{i:p}). \tag{15}$$

Under ML estimation, one can show that the variance components are given by:

$$\hat{\sigma}_{p,\text{ML}}^2 = \frac{\frac{P}{P-1}\text{MS}_p - \text{MS}_{i:p}}{I}, \quad \text{and} \quad \hat{\sigma}_{i:p,\text{ML}}^2 = \text{MS}_{i:p} \quad \text{(Searle,1997)}. \tag{16}$$

Figure 1 displays likelihood functions for nine of the 81 simulation conditions.

Bayesian Approaches. Bayesian analysis, as mentioned in the introduction, combines the likelihood with a distribution that reflects one's belief about the true nature of the variance components estimates. Specifically, prior distributions for variance components should have support on the set of nonnegative real values. EMS and ML estimates do not have such distributional constraints, and it is from this lack of constraints that inadmissible estimates arise.

Ideally, the choice of prior distribution should be invariant to the resulting posterior distribution. Three prior distributions, all of which assume independent variance

components, will be compared to assess the sensitivity of the prior to inference. These priors, are discussed below.

Each of the priors used in this study take the form of an inverse-gamma prior, $\Gamma^{-1}(\sigma^2;\alpha,\beta)$, takes the form

$$p(\sigma^2|\alpha,\beta) \propto \left(\frac{\text{SS}}{\sigma^2}\right)^{-(\alpha+1)} e^{-\beta\frac{\text{SS}}{\sigma^2}}, . \tag{17}$$

The likelihood takes the form of an χ^{-2} distribution, which is a special case of the Γ^{-1} distribution. Letting ν denote the degrees of freedom, the $\chi^{-2}(\sigma^2,\nu)$ distribution is equivalent to the $\Gamma^{-1}(\frac{\nu}{2},\frac{1}{2})$ distribution. Hence, the posterior of an inverse-chi-squared distribution, combined with an inverse-gamma prior, is also an inverse-gamma distribution:

$$p(\sigma^2_p,\sigma^2_{i:p}|\text{data}) \propto \left(\frac{\text{SS}_p}{\sigma^2_p}\right)^{-\frac{\alpha_p+2}{2}} \left(\frac{\text{SS}_{i:p}}{\sigma^2_{i:p}}\right)^{-\frac{\alpha_{i:p}+2}{2}} e^{\left[\frac{-\frac{1}{2}\text{SS}_p+(\beta_p+1/2)}{\sigma^2_p} - \frac{\frac{1}{2}\text{SS}_{i:p}+(\beta_{i:p}+1/2)}{I\sigma^2_p+\sigma^2_{i:p}}\right]}. \tag{18}$$

With some work, one can show that the MAP estimates are

$$\hat{\sigma}^2_{p,\Gamma^{-1}} = \text{SS}_p\frac{\beta_p+\frac{1}{2}}{\nu_p+1}, \quad \text{and} \quad \hat{\sigma}^2_{i:p,\Gamma^{-1}} = \text{SS}_{i:p}\frac{\beta_{i:p}+\frac{1}{2}}{\nu_{i:p}+1} \quad \text{where} \tag{19}$$

$$\nu_p = \alpha_p + \frac{P-1}{2}, \quad \text{and} \quad \nu_{i:p} = \alpha_{i:p} + \frac{P(I-1)}{2}. \tag{20}$$

Some choices for hyperprior parameters (α,β) include setting $(\alpha,\beta)=(\epsilon,\epsilon)$, for $\epsilon = 0.001$ (Spiegelhalter, Thomas, Best, & Lunn, 2004; Browne & Draper, 2006), and $(\alpha,\beta)=(-2,0)$ (Gelman et al., 2003). The choice $(\alpha,\beta)=(-1,0)$ is equivalent to a uniform prior on the interval $(0,\frac{1}{\epsilon})$ (Browne & Draper, 2006). Lastly, the choice $(\alpha,\beta)=(1,0)$ is equivalent to a posterior based on a Jeffreys' prior, $p(\sigma^2_p,\sigma^2_{i:p}) \propto (I\sigma^2_p+\sigma^2_{i:p})^{-1}(\sigma^2_{i:p})^{-1}$ (Box & Tiao, 1973).

Standardization of Sums of Squares. Even though the generating parameters, σ^2_p and $\sigma^2_{i:p}$, are constrained such that $\sigma^2_X = 100$, this constraint does not guarantee that (the observed score variance) $\hat{\sigma}^2_T = 100$. In order to ensure comparability across replications in

each condition, the sums of squares are standardized such that

$$\mathrm{SS}_p^* + \mathrm{SS}_{i:p}^* = (PI-1)\sigma_X^2, \quad \text{where} \tag{21}$$

SS_p^* and $\mathrm{SS}_{i:p}^*$ denote the standardized sums of squares,

$$\mathrm{SS}_p^* = \mathrm{SS}_p \frac{(PI-1)\sigma_X^2}{\mathrm{SS}_T}, \quad \text{and} \quad \mathrm{SS}_{i:p}^* = \mathrm{SS}_{i:p} \frac{(PI-1)\sigma_X^2}{\mathrm{SS}_T}. \tag{22}$$

Measures of Bias and Standard Error. The ML estimate for error, $\hat{\sigma}^2_{i:p,\mathrm{ML}}$, like the EMS estimators, is unbiased. This is not the case, though for the corresponding estimate for persons, $\hat{\sigma}^2_{p,\mathrm{ML}}$. The expected value of this estimate is

$$\begin{aligned} E\left(\hat{\sigma}^2_{p,\mathrm{ML}}\right) &= E\left(\frac{P-1}{P}\mathrm{MS}_p - \mathrm{MS}_{i:p}\right) \\ &= \frac{P-1}{P}\left(I\sigma_p^2 + \sigma_{i:p}^2\right) - \sigma_{i:p}^2 \\ &= \frac{IP-1}{P}\sigma_p^2 - \frac{1}{P}\sigma_{i:p}^2 \end{aligned}$$

Hence, the bias associated with $\hat{\sigma}^2_{p,\mathrm{ML}}$ is

$$\mathrm{BIAS}_{p,\mathrm{ML}} = \frac{IP-I-P}{P}\sigma_p^2 - \frac{1}{P}\sigma_{i:p}^2. \tag{23}$$

With a similar argument, one can show that, the MAP variance component estimates with the inverse-gamma prior are biased:

$$\begin{aligned} E(\sigma_p^2) &= \frac{\beta_p + \frac{1}{2}}{\nu_p + 1} E(\mathrm{SS}_p) && (24) \\ &= \frac{\beta_p + \frac{1}{2}}{\nu_p + 1}(P-1)(I\sigma_p^2 + \sigma_{i:p}^2) && (25) \\ E(\sigma_{i:p}^2) &= \frac{\beta_{i:p} + \frac{1}{2}}{\nu_{i:p} + 1} E(\mathrm{SS}_{i:p}) && (26) \\ &= \frac{\beta_{i:p} + \frac{1}{2}}{\nu_{i:p} + 1} P(I-1)\sigma_{i:p}^2 && (27) \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{BIAS}_{p,\Gamma^{-1}} &= \frac{\beta_p + \frac{1}{2}}{\nu_p + 1}(P-1)(I\sigma_p^2 + \sigma_{i:p}^2) - \sigma_p^2, \quad \text{and} && (28) \\ \mathrm{BIAS}_{i:p,\Gamma^{-1}} &= \frac{\beta_{i:p} + \frac{1}{2}}{\nu_{i:p} + 1} P(I-1)\sigma_{i:p}^2 - \sigma_{i:p}^2. && (29) \end{aligned}$$

Equations (23) through (29) are related to sample size. For large numbers of persons, the ML estimate becomes less biased. In the $i : p$ nested model $\hat{\sigma}^2_{p,\Gamma^{-1}}$ and $\hat{\sigma}^2_{i:p,\Gamma^{-1}}$ will be unbiased, respectively, for large samples of persons and for large samples of items within persons.

Combining the frequentist and Bayesian approaches, this study will estimate, for both balanced and unbalanced data, variance components using the following models:

1. EMS, Equation (11)
2. Federer-corrected EMS, Equation (12)
3. ML, Equation (16)
4. Bayes, with $\frac{\mathrm{SS}_p}{\sigma^2_p} \sim \Gamma^{-1}(-2, 0)$, and $\frac{\mathrm{SS}_{i:p}}{\sigma^2_{i:p}} \sim \Gamma^{-1}(-2, 0)$, (IG4; Browne & Draper, 2006).
5. Bayes, with $\frac{\mathrm{SS}_p}{\sigma^2_p} \sim \Gamma^{-1}(0, 0)$, and $\frac{\mathrm{SS}_{i:p}}{\sigma^2_{i:p}} \sim \Gamma^{-1}(0, 0)$, (IG5; Spiegelhalter et al., 2004).
6. Bayes, with $\frac{\mathrm{SS}_p}{\sigma^2_p} \sim \Gamma^{-1}(1, 0)$, and $\frac{\mathrm{SS}_{i:p}}{\sigma^2_{i:p}} \sim \Gamma^{-1}(1, 0)$, (Box & Tiao, 1973).

Numerical measures of bias and accuracy, in terms of recovering the generating variance components, will be obtained for each method. In order to remove effects of magnitude across conditions, relative measures of bias and standard error are computed: Let $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ denote respectively the true parameter $(\sigma^2_p, \sigma^2_{i:p})$ and its estimate $(\hat{\sigma}^2_p, \hat{\sigma}^2_{i:p})$, and let $\boldsymbol{D}$ denote the deviation

$$\boldsymbol{D} = \frac{\hat{\boldsymbol{P}} - \boldsymbol{P}}{||\boldsymbol{P}||}, \tag{30}$$

where $||\boldsymbol{P}||$ denotes the norm of $\boldsymbol{P}$. Then, the root mean square relative bias (RMSRB) and error (RMSRE) are defined, as

$$\mathrm{RMSRB} = \sqrt{||\mathrm{E}(\boldsymbol{D})||}, \quad \text{and} \quad \mathrm{RMSRE} = \sqrt{\mathrm{E}(||\boldsymbol{D}||)}, \tag{31}$$

where $E(\cdot)$ denotes the expectation operator.

Results

Balanced Data

Figures 2 and 3 display the overall RMSRB and RMSRE for each of the six estimation methods in the balanced data cases. The figures suggest no noticeable differences among the methods, when considered across all conditions, but

Insert Figures 2 and 3 here.

suggest a strong similarity between the measures of bias and error. Figure 4 displays the RMSRB and RMSRE measures across all levels of persons, items, and estimation methods (the scale of the plot permits a comparison with corresponding values from unbalanced data). One may find interesting the strong relationship between these values for large magnitudes (i.e., above 0.9).

Insert Figure 4 here.

The pearson correlation of RMSB and RMSE is 0.96. Hence, further discussion about balanced data will focus only on bias, since the results will be similar to those for error.

Figure 5 shows the RMSRB for the six estimation methods for balanced data. The figures suggests a gradually increasing relationship between numbers of persons and bias. No noticeable differences were apparent across the methods of estimation.

Insert Figure 5 here.

With respect to numbers of items for balanced data, as shown in Figure 6, bias tended to decrease with increasing numbers of items. Again, no noticeable differences among estimation methods were present.

Insert Figure 6 here.

The relationship between test reliability with bias, as shown in Figure 7, was similar to those previously discussed. Bias decreased as test reliability improved. No noticeable differences among estimation methods were found.

Insert Figure 7 here.

Unbalanced Data

Figures 8 and 9 display the overall RMSRB and RMSRE for each of the six estimation methods in the balanced data cases. Unlike the case with balanced data, noticeable differences in the distributions of bias and error were present based on estimation method. Across all conditions, the least biased methods were IG5, Box-Tiao, and EMS. The distributions of bias for the Box-Tiao and EMS methods were more similar in the sense that they tended towards slightly larger values of bias than did the distribution for IG5.

Insert Figures 8 and 9 here.

For the unbalanced data conditions, the magnitude of RMSRE values tended to be greater than those of RMSRB values. Some estimation methods, the IG5 and Box-Tiao methods in particular, had far less variability in measures of error relative to the other methods. There was a strong linear (and perhaps minimally curvilinear) relationship between the measures of bias and error, as shown in Figure 10. One remarkable difference between the scatterplots for the balanced and unbalanced conditions was the spread of the values for bias and error. There was more variability in these measures for unbalanced data relative to those for balanced data, a fact that one may notice from the histograms in Figures 8 and 9.

Insert Figure 10 here.

Again, because of the strong relationship between measures of bias and error ($r = 0.94$), further analysis will only address the former.

Figure 11 displays bias by estimation method, as a function of effective test length. The jaggedness of the relationship stems from the difference between balanced and mildly unbalanced designs. For example, a design in which four people respectively responded to 10, 10, 30, and 30 items had an effective test length of 28, and had a higher mean bias than did a balanced design where four persons each responded to 25 items. Smaller values of effective test length tended to correspond to balanced designs. The most unbalanced design corresponded to two persons, responding to 10 and 90 items, respectively. Across the range of effective test lengths, the Box-Tiao and IG5 methods tended to produce the least amount of bias.

Insert Figures 11 here.

When compared with balanced data designs, test reliabilities told a different story when explaining bias. ML was the most biased method, producing on mean RMSRB values above 1.0. Conversely, the IG5 method was the best with mean RMSRB values around 0.7. With this method, the RMSRB was lowest for tests with a reliability of 0.8, and slightly increased for tests with reliabilities of 0.9. Box-Tiao estimates tended to follow IG5 estimates for unreliable tests (i.e., tests with reliabilities below 0.5), but became more biased as reliability increased.

Insert Figures 12 here.

Conclusion and Discussion

Among the balanced data simulations, each of the estimation methods studied produced comparable measures of root mean squared relative bias and error. The distributions of bias and error were correlated highly enough that one could focus only on

analyses of bias without loss of generality. Bias tended to increase with increasing numbers of persons, but decreased with increasing numbers of items. This result is consistent with the existing literature, at least in the sense that having more items permits more informative inferences on a person' score (which is the dependent variable in any generalizability study).

Examination of the relationship between bias and test reliability suggested that Bayesian methods produced less biased variance component estimates for tests of any practical value (i.e., with reliabilities above 0.6). Federer's corrected EMS method was the least biased method of estimation for tests with reliabilities below 0.6.

Unlike the case with balanced data, there were marked differences in the distributions of bias for each of the estimation methods for unbalanced data. The $\Gamma^{-1}(-2,0)$ prior did not perform as well as expected, demonstrating relatively large bias, and was not even able to estimate measures of bias in some cases. The $\Gamma^{-1}(0,0)$ prior performed as well as the Box-Tiao prior in most cases, but produced less biased estimates in cases of marked lack of balance among items per person, as well as in cases with highly reliable tests.

Research in this area is far from sufficient. The following comments suggest directions for future work: The assumption that person and error effects are normally distributed arises mostly out of convenience. Under conditions of normality, variances follow inverse chi-squared distributions. More flexible models deserve consideration, at least to determine how realistic the normality assumption is. Bayesian nonparametric statistics (Ghosh & Ramamoorthi, 2003) allows one to incorporate uncertainty in the underlying distribution of the random effects.

The assumption of independent variance components is invalid in a fixed-variance study such as this one, and may be invalid in other studies. Some researchers, particularly in the animal science literature (e.g., Van Tassell & Van Vleck, 1996; Noguera,

Varona, Babot, & Estany, 2002; Stock, Distl, & Hoeschele, 2007), have used inverse-Wishart priors to reflect possible dependence among components. Future research may formulate a multivariate inverse-gamma distribution, which would be useful in variance component studies.

This study only examined differences among estimation methods in a one-at-time manner. Interactions, for example, between numbers of persons and items, may show differences among estimation methods.

Effective test length, at least as used in this study, was not very informative for discerning the relationship between lack of balance and relative bias. Future work should explore alternative measures of deviation from balanced designs.

Finally, it is important to consider the reason for which one chooses to estimate variance components. Interest in partitioning the variability of person scores, perhaps in the spirit of pre-work for decision studies, is certainly a valid reason. However, when the purpose is to estimate test reliability, other approaches may be more appropriate. Test reliability is a ratio of variance components, and ratio estimators do not always possess desirable characteristics. One approach around this is to derive a likelihood and posterior distribution directly in terms of reliability (Spiegelhalter, 2001), thus avoiding the need to estimate variance components.

References

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. New York: Wiley.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473-514.

Crocker, L., & Algina, J. (1986). *Classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, C. G. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Society*, *16*, 137-163.

Federer, W. T. (1968). Non-negative estimators for components of variance. *Applied Statistics*, *17*, 171-174.

Gelman, A. (2006). Prior distributions for variance component parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis*, *1*, 515-534.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (Second ed.). New York: Chapman & Hall.

Ghosh, J., & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. New York: Springer.

Hill, B. M. (1965). Inferences about variance components in the one-way model. *Journal of the American Statistical Association*, *60*, 806-825.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

Lord, F. L., & Novick, M. R. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley.

Nelder, J. A. (1954). The interpretation of negative components of variance. *Biometrika*, *41*, 544-548.

Noguera, J., Varona, L., Babot, D., & Estany, J. (2002). Multivariate analysis of litter size for multiple parities with production traits in pigs: I. Bayesian variance component estimation 1. *Journal of animal science*, *80*(10), 2540–2547.

Novick, M. R., Jackson, P. H., & Thayer, D. T. (1971). Bayesian inference and the classical test model: Reliability and true score. *Psychometrika*, *36*, 261-288.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Searle, S. R. (1997). *Linear models*. New York: Wiley.

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. New York: Wiley.

Spiegelhalter, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*, *20*, 435-452.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2004). *Winbugs version 2.0 users manual.* Available from `http://mathstat.helsinki.fi/openbugs`

Stock, K. F., Distl, O., & Hoeschele, I. (2007). Influence of priors in bayesian estimation of genetic parameters for multivariate threshold models using gibbs sampling. *Genetics, Selection, and Evolution*, *39*, 123-137.

Tiao, G. G., & Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. i. posterior distribution of variance components. *Biometrika*, *52*, 37-53.

Van Tassell, C. P., & Van Vleck, L. D. (1996). Multiple-trait gibbs sampler for animal models: flexible programs for bayesian and likelihood-based covariance component inference. *Journal of Animal Science*, *74*, 2586-2597.

Appendix

To show the relationship between variance components and test reliability for tests with a fixed-variance, begin with an expansion of terms in the relationship between person, error, and total sums of squares:

$$\begin{aligned} \mathrm{SS}_p + \mathrm{SS}_{i:p} &= \mathrm{SS}_T && (32) \\ (P-1)\mathrm{MS}_p + P(I-1)\mathrm{MS}_{i:p} &= (IP-1)\sigma_X^2 \\ (P-1)(I\sigma_p^2 + \sigma_{i:p}^2) + P(I-1)\sigma_{i:p}^2 &= (PI-1)\sigma_X^2 \\ I(P-1)\sigma_p^2 + [(P-1) + P(I-1)]\,\sigma_{i:p}^2 &= (IP-1)\sigma_X^2 \\ I(P-1)\sigma_p^2 + [(P-1) + P(I-1)]\,\sigma_{i:p}^2 &= (IP-1)\sigma_X^2 && (33) \end{aligned}$$

From Equation (33), one can substitute

$$\sigma_{i:p}^2 = \frac{1-\rho}{\rho}\sigma_p^2 \quad (34)$$

to show that

$$\begin{aligned} I(P-1)\sigma_p^2 + (IP+P-2)\frac{1-\rho}{\rho}\sigma_p^2 &= (IP-1)\sigma_X^2 \\ \left[I(P-1) + (IP+P-2)\frac{1-\rho}{\rho}\right]\sigma_p^2 &= (IP-1)\,,\sigma_X^2 && (35) \end{aligned}$$

from which it follows that

$$\sigma_p^2 = \frac{(IP-1)}{I(P-1) + (IP+P-2)\frac{1-\rho}{\rho}}\,\sigma_X^2. \quad (36)$$

Solving for $\sigma_{i:p}^2$ requires substituting

$$\sigma_p^2 = \frac{\rho}{1-\rho}\sigma_{i:p}^2 \quad (37)$$

in Equation (33), which results in

$$\begin{aligned} I(P-1)\frac{\rho}{1-\rho}\sigma_{i:p}^2 + (IP+P-2)\sigma_{i:p}^2 &= (IP-1)\sigma_X^2 \\ \left[I(P-1)\frac{\rho}{1-\rho} + (IP+P-2)\right]\sigma_{i:p}^2 &= (IP-1)\sigma_X^2 \\ \sigma_{i:p}^2 &= \frac{(IP-1)}{I(P-1)\frac{\rho}{1-\rho} + (IP+P-2)}\,\sigma_X^2. && (38) \end{aligned}$$

Table 1

Expected mean squares for the one-way ANOVA model.

Source	d.f.	E.M.S.
Persons	$\nu_p = P - 1$	$E(\mathrm{MS}_p) = \sigma^2_{i:p} + K\sigma^2_p$
Error	$\nu_{pk} = P(K-1)$	$E(\mathrm{MS}_{i:p}) = \sigma^2_{i:p}$

Table 2

True variance components for generating random values for $P = 25$.

P	I	ρ	σ^2_p	$\sigma^2_{i:p}$	P	I	ρ	σ^2_p	$\sigma^2_{i:p}$	P	I	ρ	σ^2_p	$\sigma^2_{i:p}$
25	10	0.10	9.23	83.09	25	50	0.10	9.87	88.81	25	100	0.10	9.95	89.58
25	10	0.20	18.69	74.77	25	50	0.20	19.85	79.40	25	100	0.20	20.00	80.02
25	10	0.30	28.39	66.25	25	50	0.30	29.95	69.88	25	100	0.30	30.16	70.36
25	10	0.40	38.34	57.51	25	50	0.40	40.17	60.25	25	100	0.40	40.41	60.61
25	10	0.50	48.54	48.54	25	50	0.50	50.51	50.51	25	100	0.50	50.76	50.76
25	10	0.60	59.00	39.34	25	50	0.60	60.97	40.64	25	100	0.60	61.22	40.81
25	10	0.70	69.75	29.89	25	50	0.70	71.55	30.67	25	100	0.70	71.78	30.76
25	10	0.80	80.78	20.19	25	50	0.80	82.27	20.57	25	100	0.80	82.45	20.61
25	10	0.85	86.41	15.25	25	50	0.85	87.67	15.47	25	100	0.85	87.83	15.50
25	10	0.90	92.11	10.23	25	50	0.90	93.11	10.35	25	100	0.90	93.23	10.36
25	10	0.95	97.89	5.15	25	50	0.95	98.58	5.19	25	100	0.95	98.67	5.19

Table 3

True variance components for generating random values for $P = 100$.

P	I	ρ	σ^2_p	$\sigma^2_{i:p}$	P	I	ρ	σ^2_p	$\sigma^2_{i:p}$	P	I	ρ	σ^2_p	$\sigma^2_{i:p}$
100	10	0.10	9.19	82.70	100	50	0.10	9.83	88.51	100	100	0.10	9.92	89.29
100	10	0.20	18.56	74.25	100	50	0.20	19.73	78.90	100	100	0.20	19.88	79.53
100	10	0.30	28.13	65.62	100	50	0.30	29.68	69.24	100	100	0.30	29.88	69.72
100	10	0.40	37.88	56.83	100	50	0.40	39.68	59.53	100	100	0.40	39.92	59.88
100	10	0.50	47.84	47.84	100	50	0.50	49.75	49.75	100	100	0.50	50.00	50.00
100	10	0.60	58.01	38.68	100	50	0.60	59.88	39.92	100	100	0.60	60.12	40.08
100	10	0.70	68.40	29.31	100	50	0.70	70.06	30.03	100	100	0.70	70.28	30.12
100	10	0.80	79.00	19.75	100	50	0.80	80.31	20.08	100	100	0.80	80.48	20.12
100	10	0.85	84.39	14.89	100	50	0.85	85.46	15.08	100	100	0.85	85.59	15.10
100	10	0.90	89.84	9.98	100	50	0.90	90.62	10.07	100	100	0.90	90.72	10.08
100	10	0.95	95.34	5.02	100	50	0.95	95.80	5.04	100	100	0.95	95.85	5.04

Table 4

True variance components for generating random values for $P = 1000$.

P	I	ρ	σ^2_p	$\sigma^2_{i:p}$	P	I	ρ	σ^2_p	$\sigma^2_{i:p}$	P	I	ρ	σ^2_p	$\sigma^2_{i:p}$
1000	10	0.10	9.18	82.58	1000	50	0.10	9.82	88.42	1000	100	0.10	9.91	89.21
1000	10	0.20	18.52	74.09	1000	50	0.20	19.69	78.76	1000	100	0.20	19.85	79.38
1000	10	0.30	28.05	65.44	1000	50	0.30	29.59	69.05	1000	100	0.30	29.80	69.53
1000	10	0.40	37.75	56.63	1000	50	0.40	39.54	59.31	1000	100	0.40	39.78	59.67
1000	10	0.50	47.64	47.64	1000	50	0.50	49.53	49.53	1000	100	0.50	49.78	49.78
1000	10	0.60	57.72	38.48	1000	50	0.60	59.56	39.71	1000	100	0.60	59.80	39.86
1000	10	0.70	68.00	29.14	1000	50	0.70	69.63	29.84	1000	100	0.70	69.84	29.93
1000	10	0.80	78.49	19.62	1000	50	0.80	79.74	19.94	1000	100	0.80	79.90	19.98
1000	10	0.85	83.81	14.79	1000	50	0.85	84.82	14.97	1000	100	0.85	84.94	14.99
1000	10	0.90	89.18	9.91	1000	50	0.90	89.90	9.99	1000	100	0.90	89.99	10.00
1000	10	0.95	94.61	4.98	1000	50	0.95	94.99	5.00	1000	100	0.95	95.04	5.00

Table 5

Proportions of and variance components for two unbalanced groups.

N_1	N_2	I_{eff}	ρ	σ^2_p	$\sigma^2_{i:p}$
10	90	82	0.1	10.6	95.3
10	90	82	0.2	22.3	89.3
10	90	82	0.3	35.4	82.7
10	90	82	0.4	50.2	75.2
10	90	82	0.5	66.8	66.8
10	90	82	0.6	85.8	57.2
10	90	82	0.7	107.6	46.1
10	90	82	0.8	133.1	33.3
10	90	82	0.9	163	18.1
50	50	50	0.1	10.6	95.6
50	50	50	0.2	22.4	89.6
50	50	50	0.3	35.5	82.9
50	50	50	0.4	50.3	75.4
50	50	50	0.5	66.9	66.9
50	50	50	0.6	85.8	57.2
50	50	50	0.7	107.6	46.1
50	50	50	0.8	132.9	33.2
50	50	50	0.9	162.6	18.1

Table 6

Proportions of and variance components for three unbalanced groups.

N_1	N_2	N_3	I_{eff}	ρ	σ_p^2	$\sigma_{i:p}^2$
10	45	45	41.5	0.1	10.4	93.8
10	45	45	41.5	0.2	21.6	86.2
10	45	45	41.5	0.3	33.5	78.1
10	45	45	41.5	0.4	46.3	69.4
10	45	45	41.5	0.5	60.1	60.1
10	45	45	41.5	0.6	75.0	50.0
10	45	45	41.5	0.7	91.1	39.1
10	45	45	41.5	0.8	108.7	27.2
10	45	45	41.5	0.9	127.8	14.2
10	10	80	66.0	0.1	10.4	93.5
10	10	80	66.0	0.2	21.5	86.0
10	10	80	66.0	0.3	33.4	78.0
10	10	80	66.0	0.4	46.2	69.4
10	10	80	66.0	0.5	60.1	60.1
10	10	80	66.0	0.6	75.0	50.0
10	10	80	66.0	0.7	91.2	39.1
10	10	80	66.0	0.8	108.8	27.2
10	10	80	66.0	0.9	128.1	14.2
33	33	34	33.3	0.1	10.4	93.9
33	33	34	33.3	0.2	21.6	86.3
33	33	34	33.3	0.3	33.5	78.2
33	33	34	33.3	0.4	46.3	69.5
33	33	34	33.3	0.5	60.1	60.1
33	33	34	33.3	0.6	75.0	50.0
33	33	34	33.3	0.7	91.1	39.0
33	33	34	33.3	0.8	108.6	27.1
33	33	34	33.3	0.9	127.7	14.2

Table 7

Proportions of and variance components for four unbalanced groups.

N_1	N_2	N_3	N_4	I_{eff}	ρ	σ_p^2	$\sigma_{i:p}^2$	N_1	N_2	N_3	N_4	I_{eff}	ρ	σ_p^2	$\sigma_{i:p}^2$
25	25	25	25	25	0.1	10.3	93.1	10	10	40	40	34	0.1	10.3	92.9
25	25	25	25	25	0.2	21.2	84.8	10	10	40	40	34	0.2	21.2	84.6
25	25	25	25	25	0.3	32.6	76.1	10	10	40	40	34	0.3	32.6	76.0
25	25	25	25	25	0.4	44.6	66.9	10	10	40	40	34	0.4	44.6	66.8
25	25	25	25	25	0.5	57.2	57.2	10	10	40	40	34	0.5	57.2	57.2
25	25	25	25	25	0.6	70.5	47.0	10	10	40	40	34	0.6	70.6	47.0
25	25	25	25	25	0.7	84.6	36.3	10	10	40	40	34	0.7	84.7	36.3
25	25	25	25	25	0.8	99.5	24.9	10	10	40	40	34	0.8	99.6	24.9
25	25	25	25	25	0.9	115.3	12.8	10	10	40	40	34	0.9	115.5	12. 8
10	30	30	30	28	0.1	10.3	93.0	10	10	10	70	52	0.1	10.3	92.7
10	30	30	30	28	0.2	21.2	84.7	10	10	10	70	52	0.2	21.1	84.5
10	30	30	30	28	0.3	32.6	76.0	10	10	10	70	52	0.3	32.5	75.9
10	30	30	30	28	0.4	44.6	66.9	10	10	10	70	52	0.4	44.5	66.8
10	30	30	30	28	0.5	57.2	57.2	10	10	10	70	52	0.5	57.2	57.2
10	30	30	30	28	0.6	70.6	47.0	10	10	10	70	52	0.6	70.6	47.0
10	30	30	30	28	0.7	84.6	36.3	10	10	10	70	52	0.7	84.7	36.3
10	30	30	30	28	0.8	99.6	24.9	10	10	10	70	52	0.8	99.8	24.9
10	30	30	30	28	0.9	115.4	12.8	10	10	10	70	52	0.9	115.7	12.9

Figure Captions

Figure 1. Likelihood functions based on generating variance components for $P = \{25, 100, 1000\}$, $I = \{10, 50, 100\}$, and $\rho = \{0.10, 0.5, 0.9\}$.

Figure 2. RMSB for balanced data, by estimation method.

Figure 3. RMSE for balanced data, by estimation method.

Figure 4. RMSB vs. RMSE for balanced data, across all conditions.

Figure 5. RMSB by estimation method, as a function of numbers of persons.

Figure 6. RMSB by estimation method, as a function of numbers of items.

Figure 7. RMSB by estimation method, as a function of test reliability.

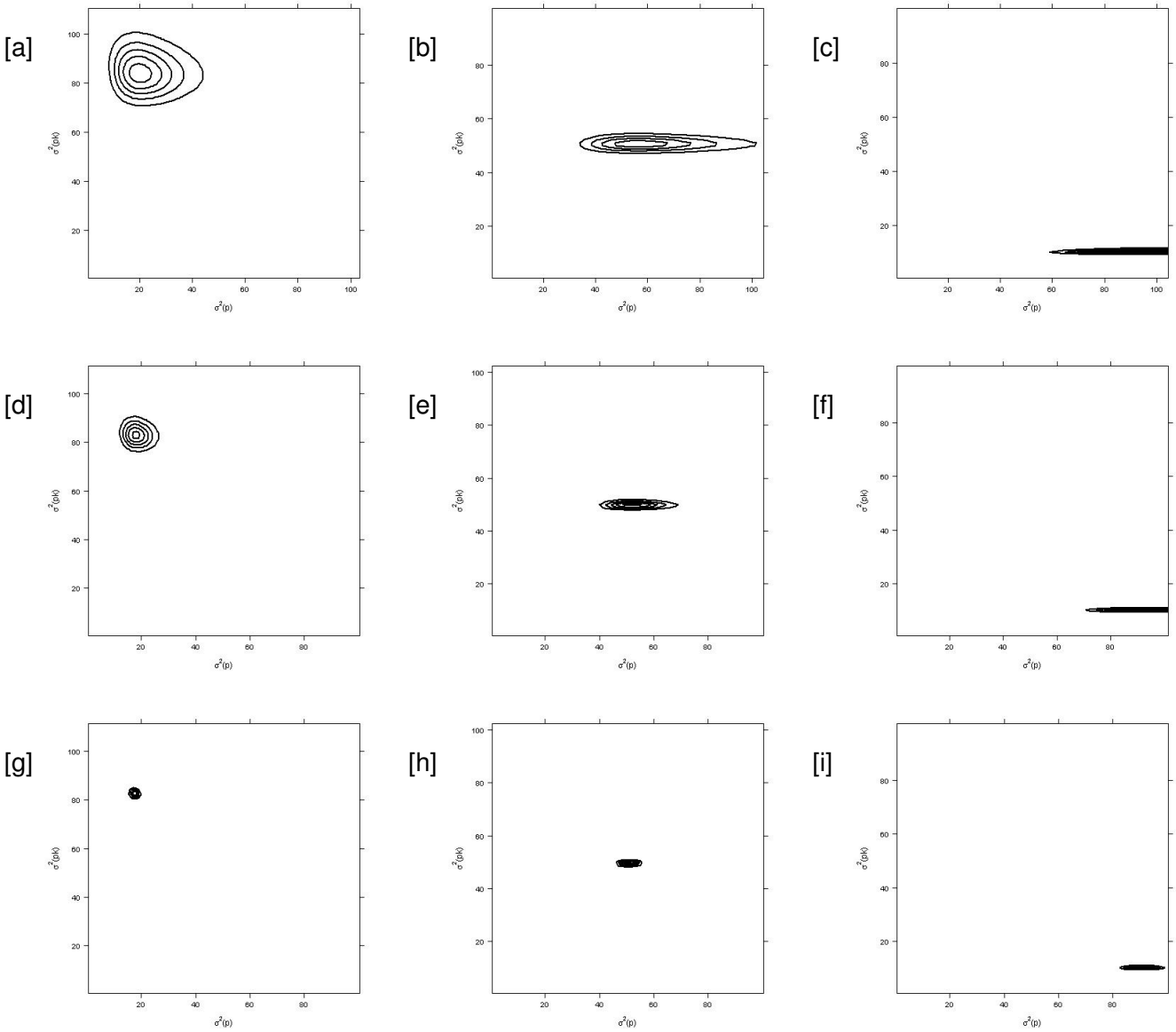
Figure 8. RMSB for unbalanced data, by estimation method.

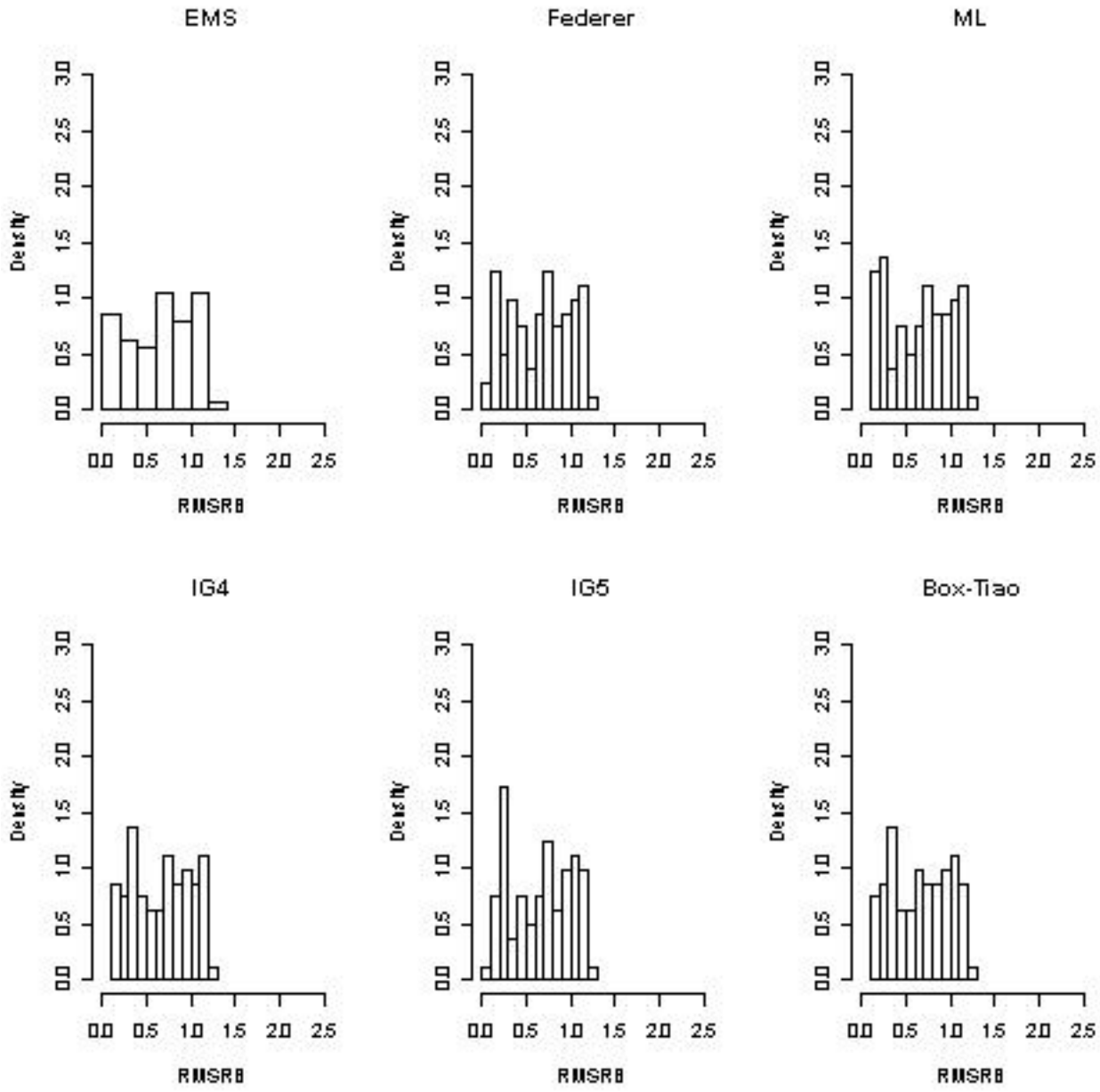
Figure 9. RMSE for unbalanced data, by estimation method.

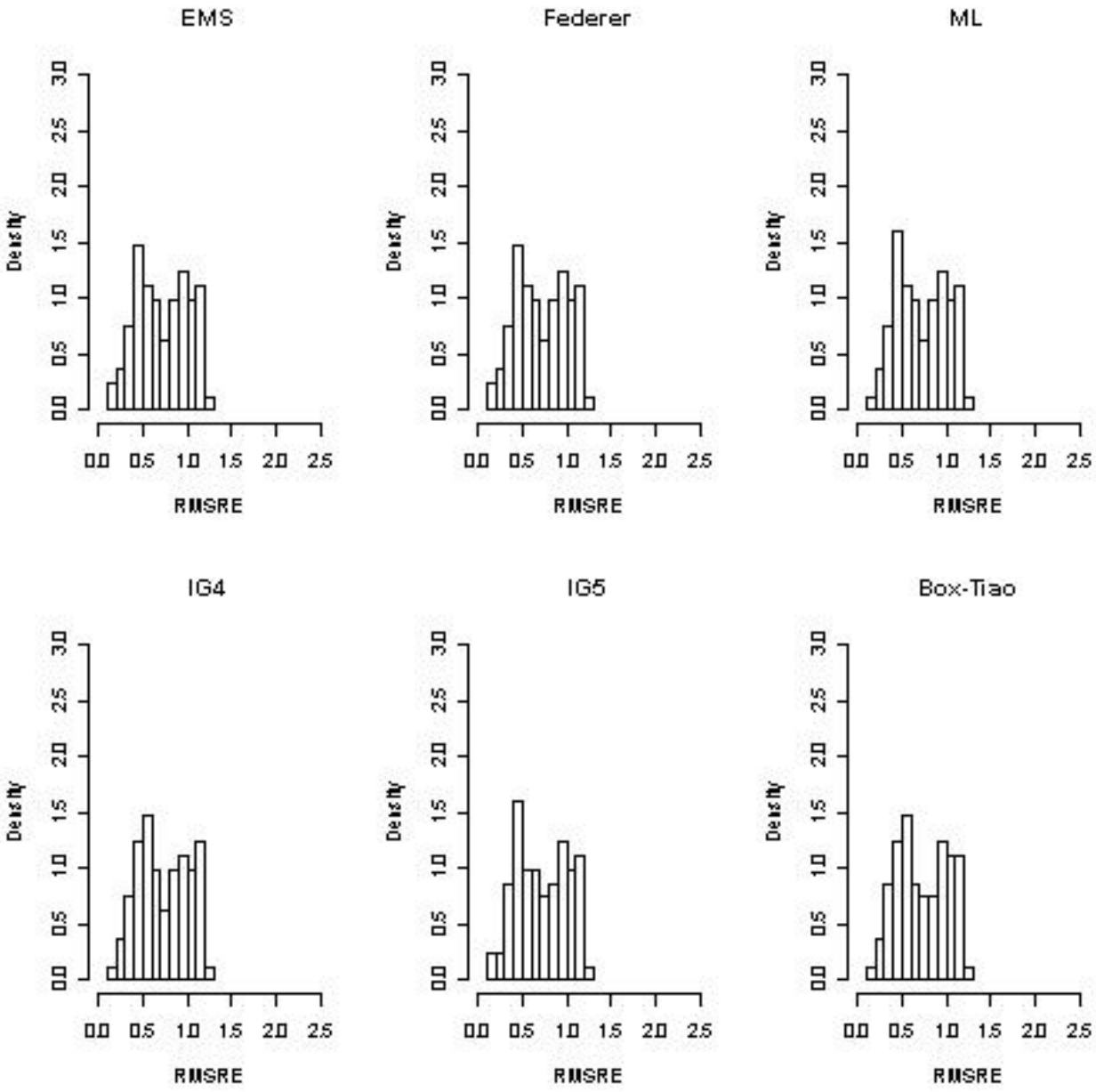
Figure 10. RMSB vs. RMSE, for unbalanced data, by estimation method.

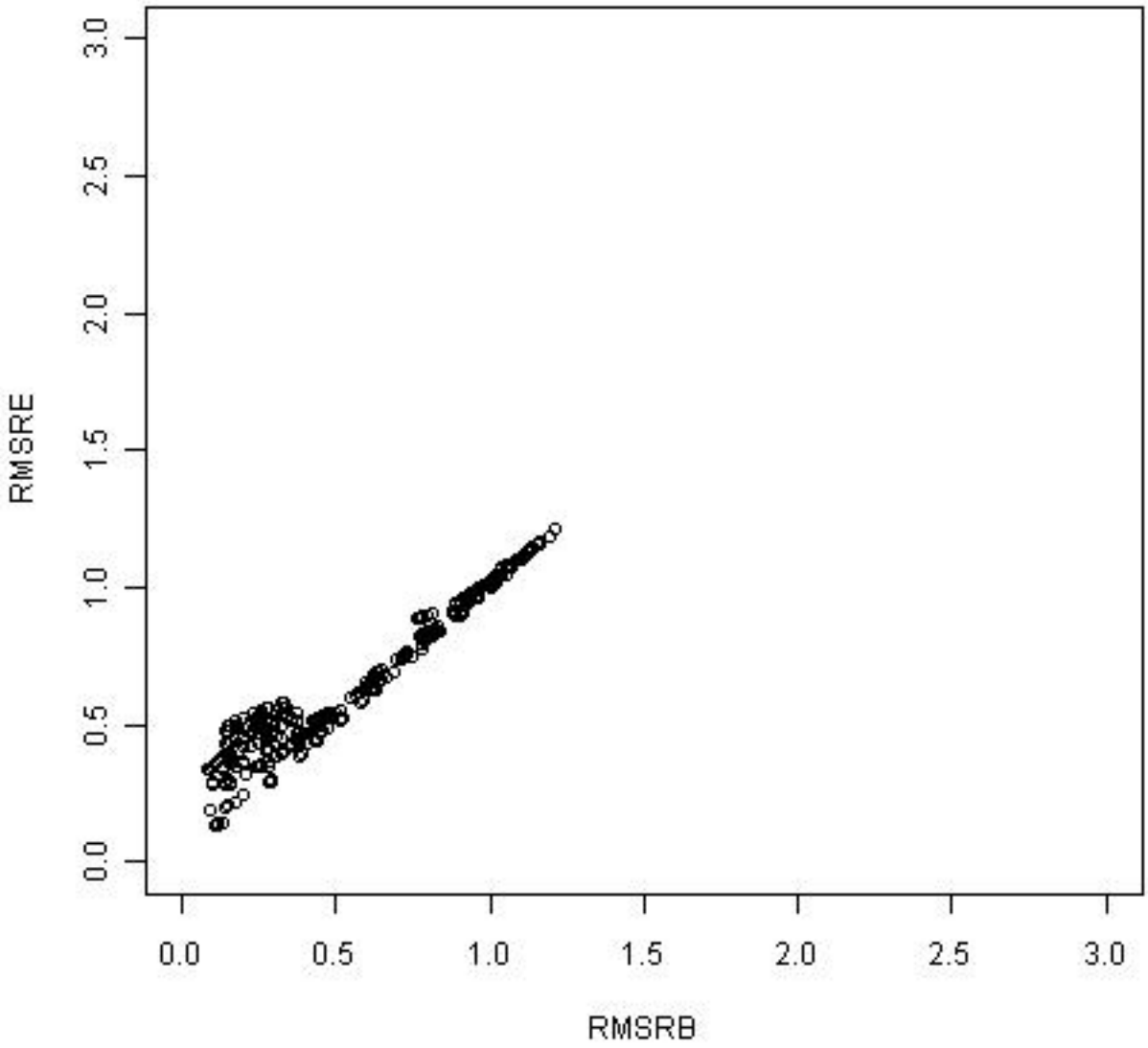
Figure 11. RMSB, for unbalanced data, by estimation method, as function of effective test length.

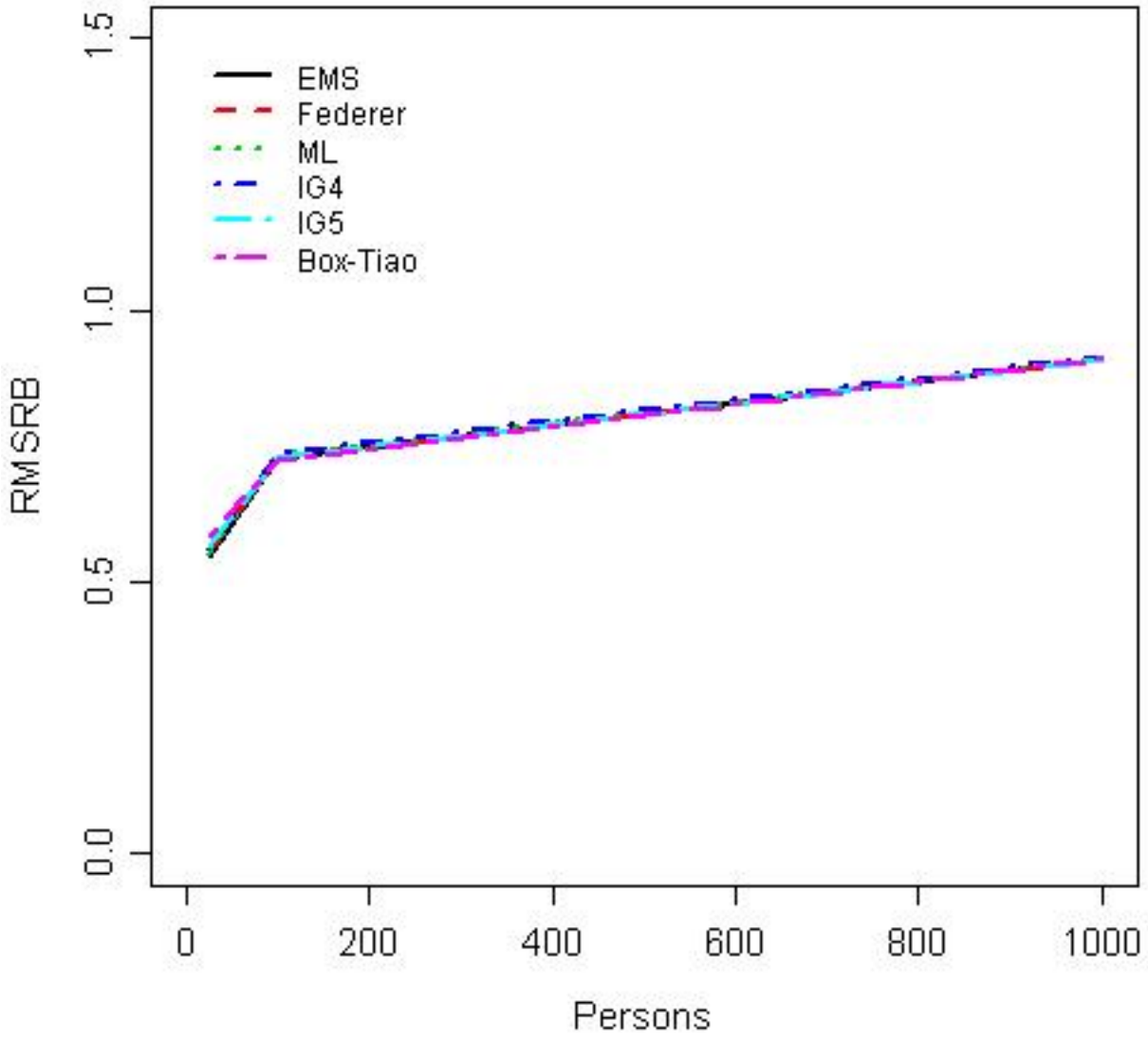
Figure 12. RMSB, for unbalanced data, by estimation method, as function of test reliability.

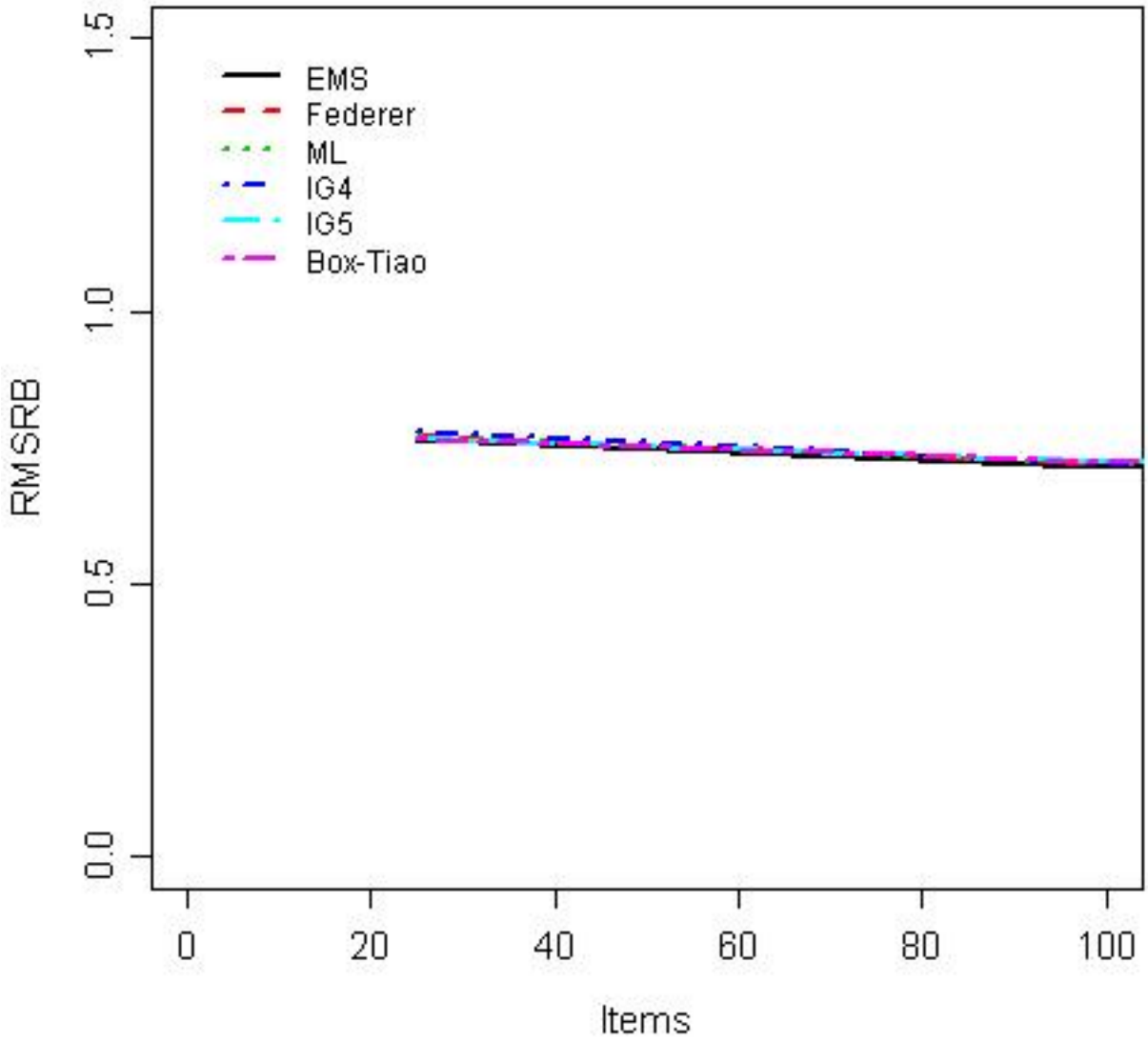

[a]
[b]
[c]
[d]
[e]
[f]
[g]
[h]
[i]
σ²(gk)
σ²(p)

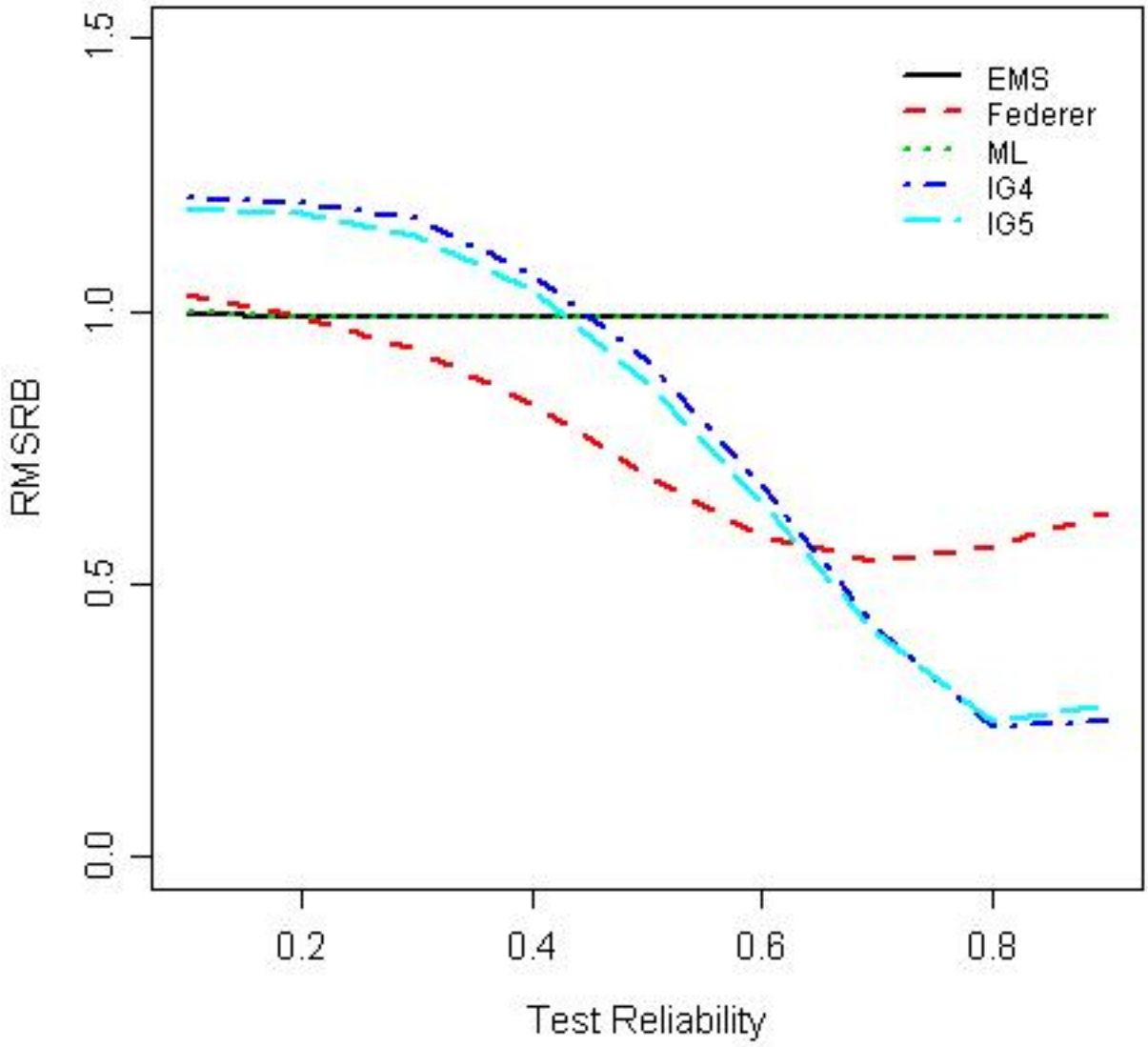

EMS
Federer
ML
IG4
IG5
Box-Tiao
Density
RMSRB
0.0
0.5
1.0
1.5
2.0
2.5
3.0

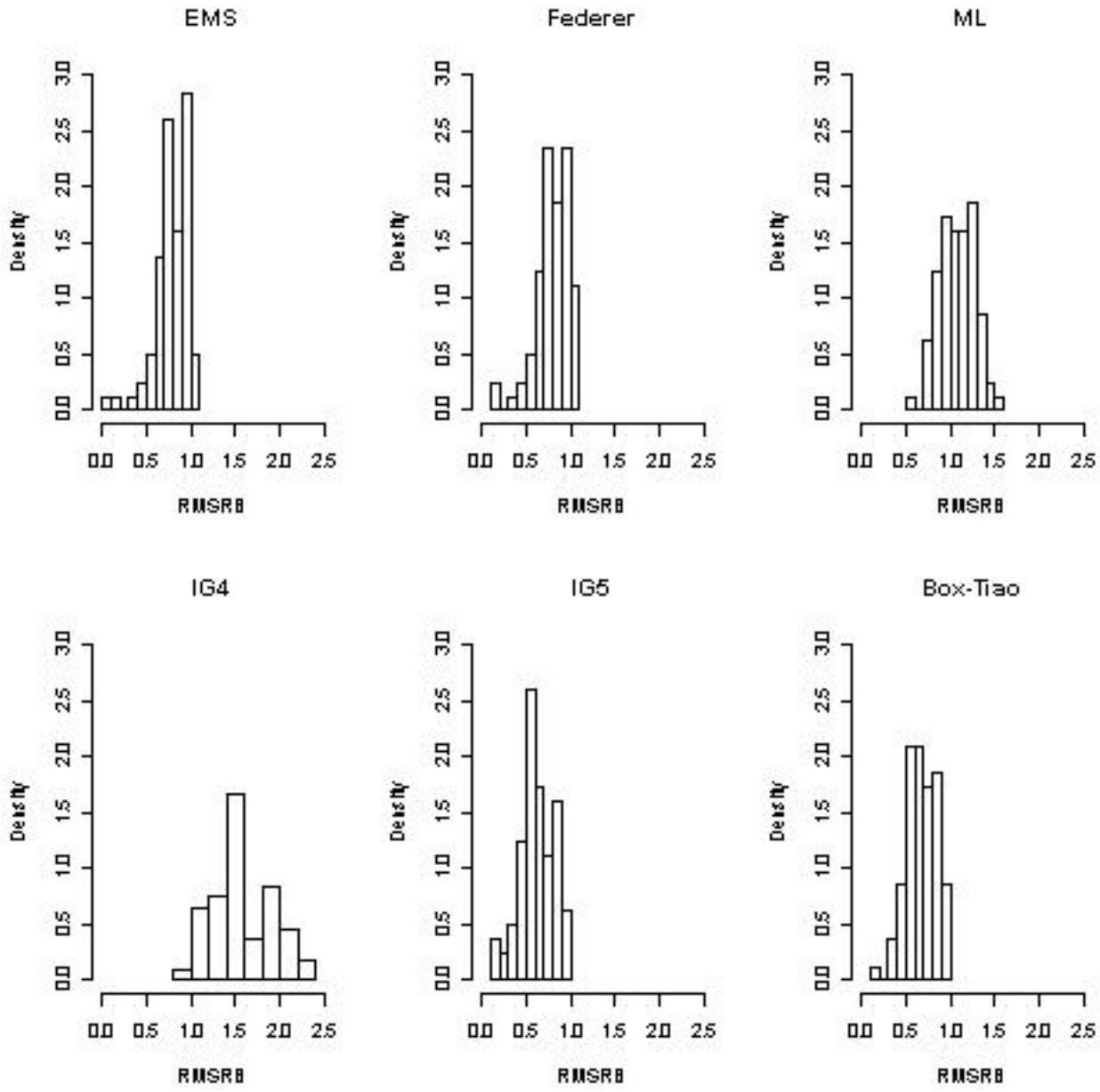

EMS
Federer
ML
IG4
IG5
Box-Tiao
Density
RMSRE
0.0
0.5
1.0
1.5
2.0
2.5
3.0

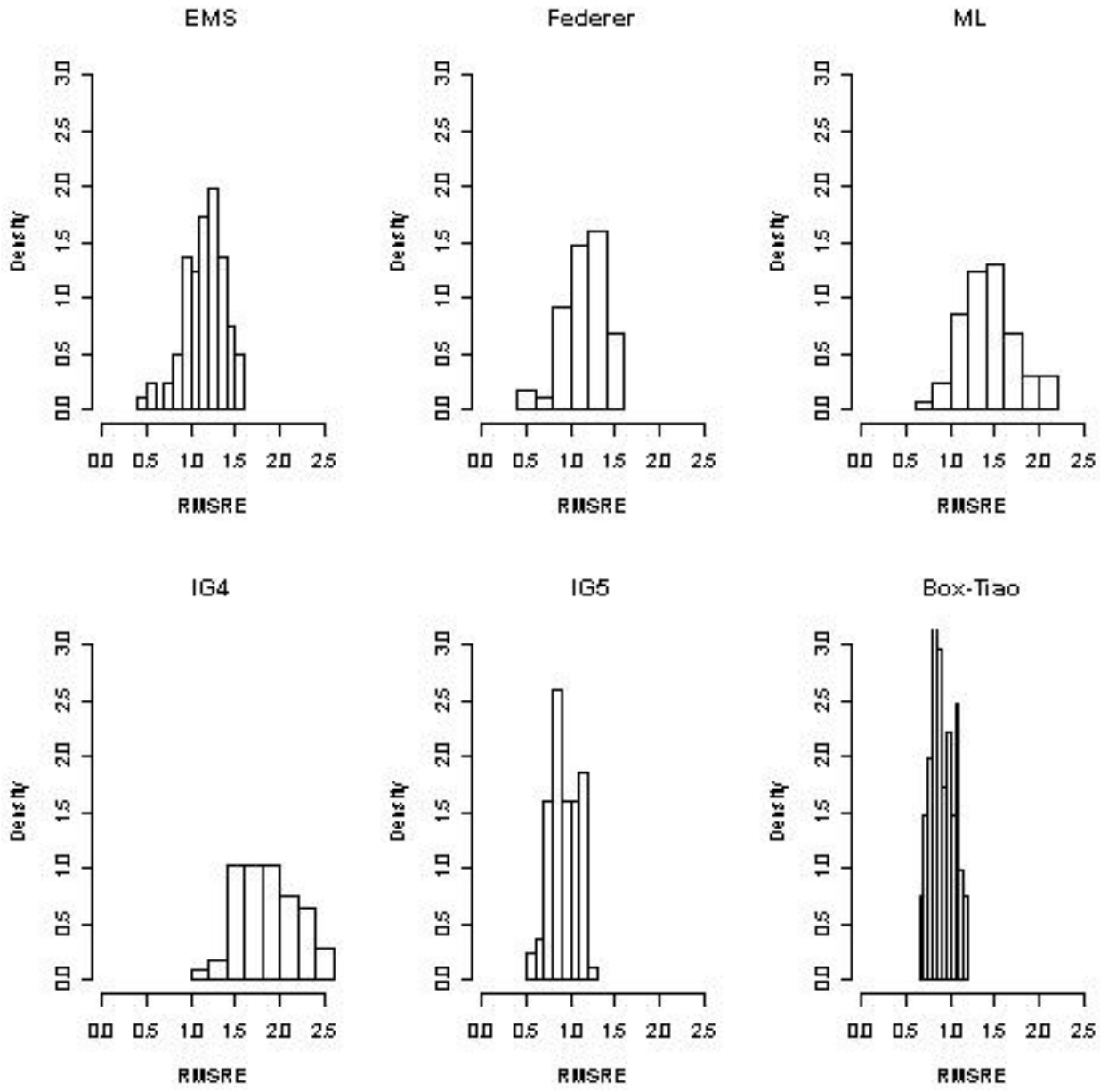

RMSRE
RMSRB
0.0
0.5
1.0
1.5
2.0
2.5
3.0

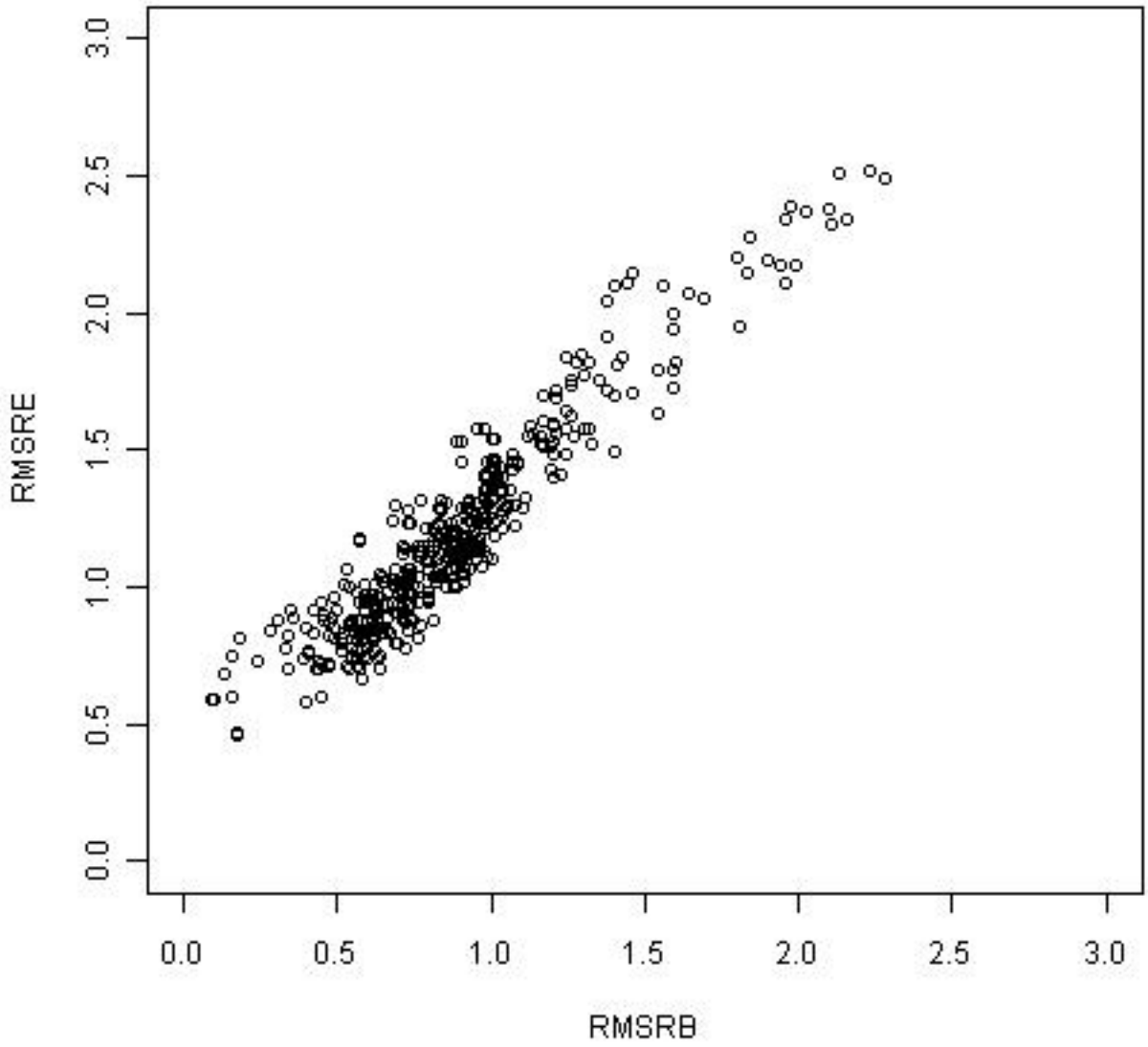

EMS
Federer
ML
IG4
IG5
Box-Tiao
RMSRB
0.0
0.5
1.0
1.5
0
200
400
600
800
1000
Persons

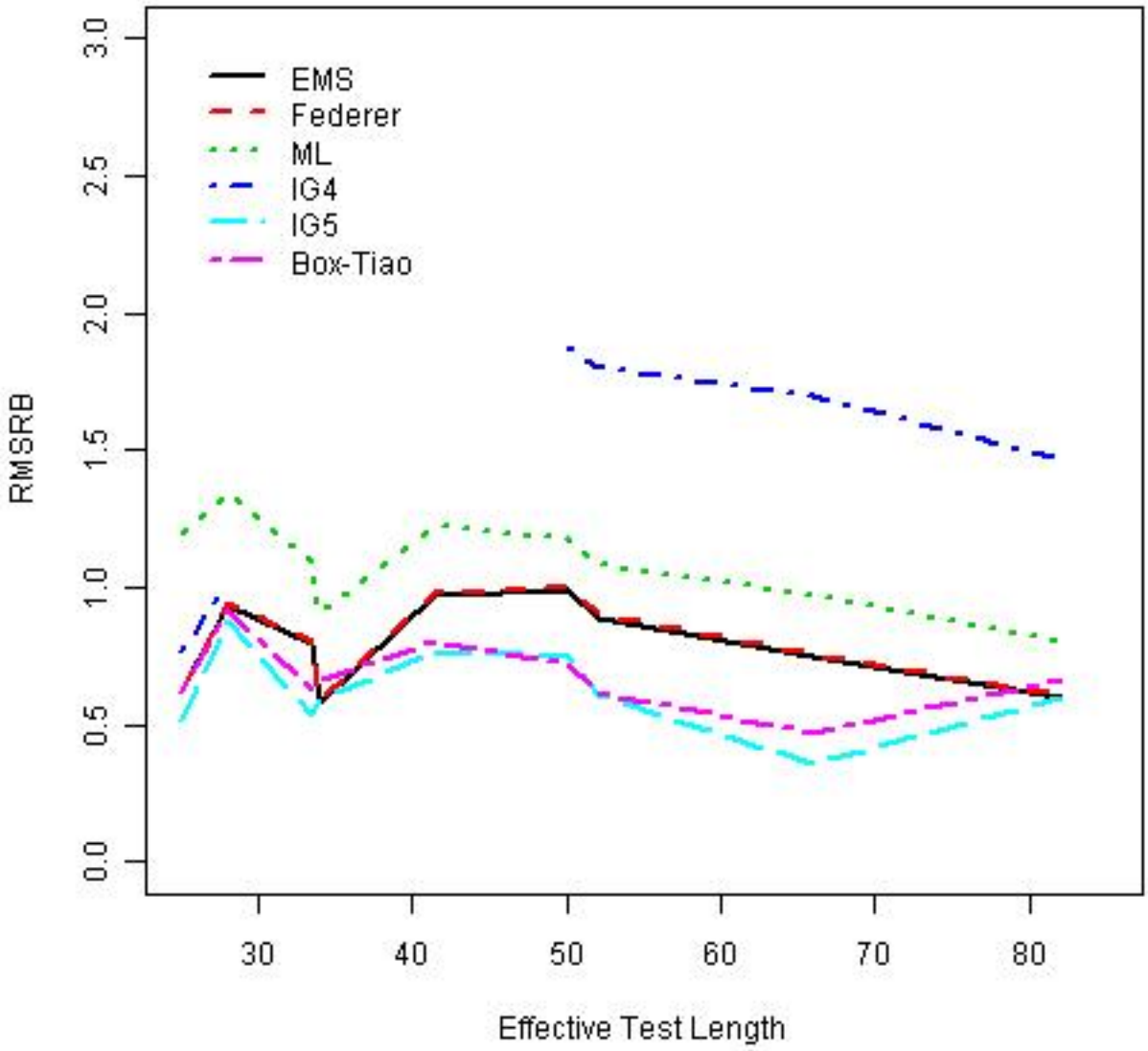

EMS
Federer
ML
IG4
IG5
Box-Tiao
RMSRB
0.0
0.5
1.0
1.5
0
20
40
60
80
100
Items

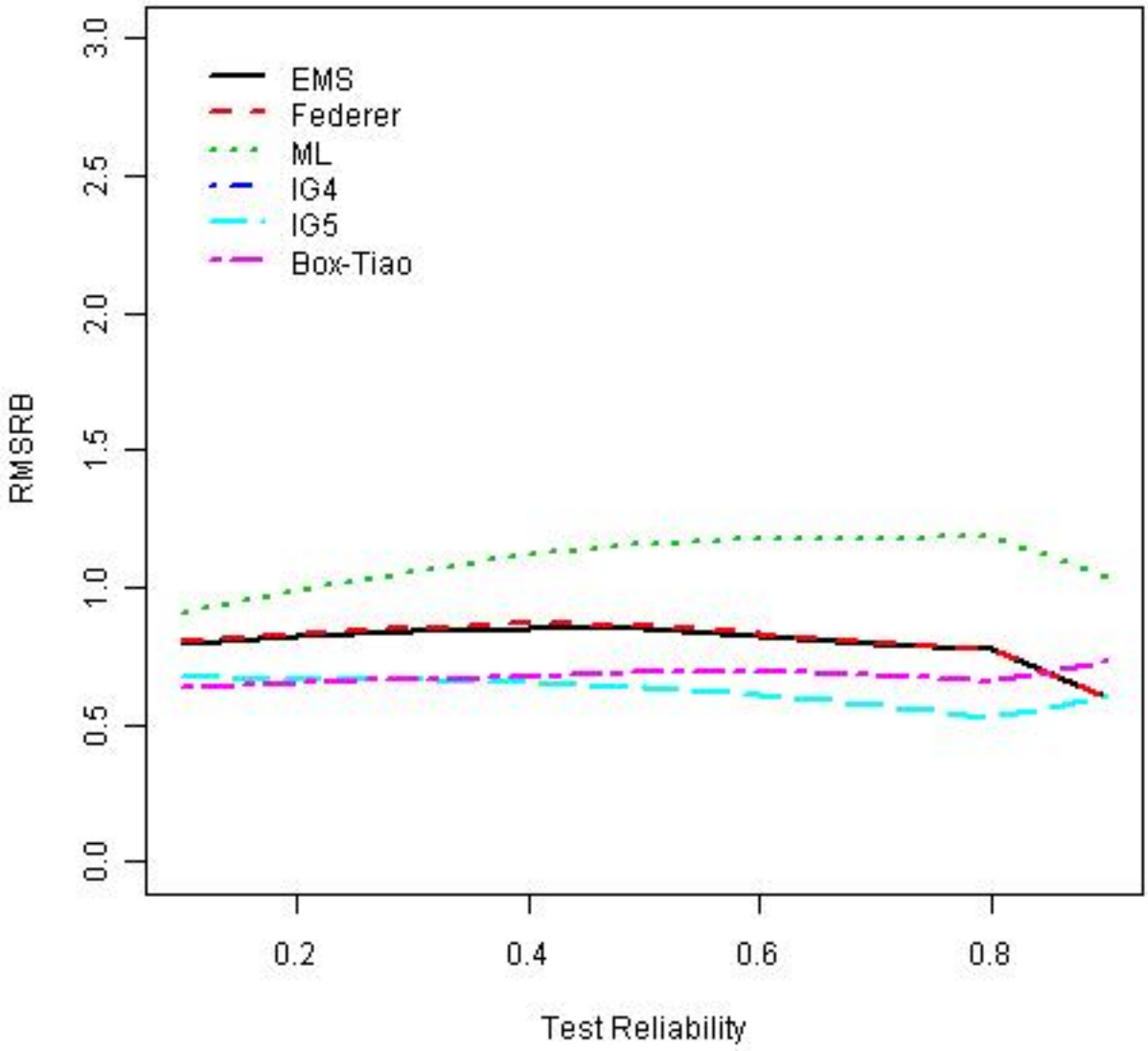

EMS
Federer
ML
IG4
IG5
RMSRB
0.0
0.5
1.0
1.5
0.2
0.4
0.6
0.8
Test Reliability


EMS
Federer
ML
IG4
IG5
Box-Tiao
Density
RMSR8
0.0
0.5
1.0
1.5
2.0
2.5
3.0


EMS
Federer
ML
IG4
IG5
Box-Tiao
Density
RMSRE
0.0
0.5
1.0
1.5
2.0
2.5
3.0


RMSRE
RMSRB
0.0
0.5
1.0
1.5
2.0
2.5
3.0


EMS
Federer
ML
IG4
IG5
Box-Tiao
RMSRB
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
40
50
60
70
80
Effective Test Length


EMS
Federer
ML
IG4
IG5
Box-Tiao
RMSRB
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8
Test Reliability